UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

September 19, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES V-2

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich DK
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Michael Starsinic
Variance Estimation Branch

Subject: Accuracy and Coverage Evaluation Survey: Overview of Census 2000 A.C.E. Variance Estimation - Theory and Implementation

The purpose of this memorandum is to offer a general overview of the entire Accuracy and Coverage Evaluation (A.C.E.) variance estimation procedures. This includes the theory behind the methodology, which is fully explored in Kim, Navarro, and Fuller [1]. The operational implementation of the methodology is described in extensive detail in DSSD Census 2000 Procedures and Operations Memorandum V-1, "Computer Specifications for Variance Estimation for Census 2000" [2]. Any questions regarding this overview should be directed to Michael Starsinic at (301) 457-1638.

cc: Variance Estimation Staff

Overview of Census 2000 A.C.E. Variance Estimation: Theory & Implementation

Introduction

Several factors combined to make the variance estimation for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) survey a very different operation than the variance estimation which was done for the Census 2000 Dress Rehearsal in Sacramento, CA and the Menominee American Indian Reservation in Wisconsin. As a result of the Supreme Court's decision to disallow the use of statistically adjusted estimates for Congressional reapportionment, the proposed sampling and estimation procedures were radically changed. First, sampling for non-response follow-up (NRFU) was eliminated. Follow-up was on 100% of non-responding housing units, as it was in the 1990 census, but unlike the methodology used in Sacramento during the Dress Rehearsal. This removed one of the two primary components of the measured variance from the Dress Rehearsal.

The second adjustment was made to the sampling scheme, which initially was designed to be of 750,000 housing units. Now, for the A.C.E., only 300,000 housing units were needed. The process to produce the sample of 750,000 was already in motion, and it was decided to continue with the operation and later subsample to bring the expected size down to 300,000. The later sampling operations of A.C.E. Reduction, Small Block Cluster Subsampling, and Large Block Cluster Subsampling would reduce the sample to approximately 300,000.

A further wrinkle was the inclusion of Targeted Extended Search. In the 1990 census, all block clusters selected in the Post-Enumeration Survey (PES) had the block clusters that surrounded them searched for possible additional matches and duplicates. For Census 2000, a sample of 20% of block clusters was selected to have their search areas extended. Half of the 20% are block clusters with high numbers of P-sample nonmatches or E-sample geocoding errors. These block clusters will be selected with certainty. The remaining 10% will come from a systematic sample of block clusters which contain at least one P-sample nonmatch or E-sample geocoding error.

The combination of the multiphase selection of the A.C.E. sample and the implementation of Targeted Extended Search made the Dress Rehearsal methodology of a stratified jackknife unusable. Specifically, the A.C.E. sample was considered a three phase sample - the initial Listing sample as the first phase; A.C.E. Reduction and Small Block Cluster Subsampling as the second phase; and TES as the third phase. Multiphase sampling differs from multistage in the following way: in a multistage design, the information needed to draw all stages of the sample is known before the sampling begins; in a multiphase design, the information needed to draw the n^{th} phase of the sample is unobtainable until the $n\text{-}1^{st}$ phase of the sample is completed. Thus, it was required to develop a new methodology to compute variance estimates.

New Methodology

Our goal is to obtain a variance estimator for the Dual System Estimator (DSE), which has the following form:

$$D\hat{S}E = (C - II)\left(\frac{CE}{N_e}\right)\left(\frac{N_n + N_i}{M_n + \left(\frac{M_o}{N_o}\right)N_i}\right) \tag{1}$$

where :

C = unweighted census count
II = count of not-data-defined and wholly imputed persons
CE = estimated number of A.C.E. E-Sample correct enumerations
N_e = estimated number of A.C.E. E-Sample persons
N_n = estimated number of A.C.E. P-Sample nonmovers
N_i = estimated number of A.C.E. P-Sample inmovers
N_o = estimated number of A.C.E. P-Sample outmovers
M_n = estimated number of A.C.E. P-Sample nonmover matches
M_o = estimated number of A.C.E. P-Sample outmover matches

The DSE is computed separately for each post-stratum. The national corrected population estimate is computed as:

$$\hat{T}_{US} = \sum_{PS} D\hat{S}E_{PS} \tag{2}$$

There is no closed-form solution for the variance estimator, and the Taylor linearization variance estimator is very complex. That leaves the replication methodology as out best bet to find a variance estimator. For a general estimator,

$$T_y = \sum_i w_i y_i \tag{3}$$

the replicate estimator is

$$T_y^{(k)} = \sum_i w_i^{(k)} y_i \tag{4}$$

where y_i is the characteristic of interest, and $w_i^{(k)}$ is the replicate weight, which differs from the original weight in a prespecified subset of the observations. With these replicate estimators, a variance estimator can be constructed:

$$\hat{Var}(T_y) = \sum_j c_j (T_y^{(k)} - T_y)^2 \tag{5}$$

Before we continue, we must set down some specific notation. Let w_i be the first phase sampling weight, and let y_i be the cluster-level total of any of the seven estimated components of the DSE (CE, N_n, etc.). Let A and A_2 indicate the first and second phase samples, respectively. Let $x_{ig}=1$ if unit i is in "group" (second phase stratum) g and zero otherwise. Let n_h be the number of units selected in first-phase stratum h. Let n_g be the number of units in stratum h that are also in group g, and let r_g be the number of the n_g units selected in the second phase.

For two-phase stratified sampling, there are two different point estimators, the Double Expansion Estimator (DEE)

$$\text{DEE} = \sum_{g} \sum_{i \in A_2} \frac{n_g}{r_g} w_i x_{ig} y_i \tag{6}$$

and the Reweighted Expansion Estimator (REE)

$$\text{REE} = \sum_{g} \sum_{i \in A_2} \left(\frac{\sum_{i \in A} w_i x_{ig}}{\sum_{i \in A_2} w_i x_{ig}} \right) w_i x_{ig} y_i \tag{7}$$

There is an established result by Rao & Shao (*Biometrika*, 1992) which gives a replicate variance estimator for the REE under two-phase stratified sampling. Unfortunately, all the individual components of the DSE, such as N_e, the number of E-Sample people, are DEE's. Taking a closer look at the DEE, however, gives a hint as to how to proceed.

$$n_g = \sum_{i \in A} x_{ig}, \quad r_g = \sum_{i \in A_2} x_{ig}$$

$$\text{DEE} = \sum_{g} \sum_{i \in A_2} \frac{n_g}{r_g} w_i x_{ig} y_i = \sum_{g} \sum_{i \in A_2} \left(\frac{\sum_{j \in A} x_{jg}}{\sum_{j \in A_2} x_{jg}} \right) w_i x_{ig} y_i = \sum_{g} \sum_{i \in A_2} \left(\frac{\sum_{j \in A} w_j x_{jg} w_j^{-1}}{\sum_{j \in A_2} w_j x_{jg} w_j^{-1}} \right) w_i x_{ig} y_i \tag{8}$$

We have just rewritten the DEE in a form which is quite similar to the REE. This suggests the following generalization:

$$T_{y2} = \sum_{i \in A_2} \alpha_i y_i, \text{ where}$$

$$\alpha_i = \sum_{g} \left(\frac{\sum_{j \in A} w_j x_{jg} q_j}{\sum_{j \in A_2} w_j x_{jg} q_j} \right) w_i x_{ig} \tag{9}$$

and where $q_j = 1$ for the REE and w_i^{-1} for the DEE.

Replicates are then naturally written as:

$$T_{y\cdot 2}^{(k)} = \sum_{i \in A_2} \alpha_i^{(k)} y_i, \text{ where}$$

$$\alpha_i^{(k)} = \sum_g \left(\frac{\sum_{j \in A} w_j^{(k)} x_{jg} q_j}{\sum_{j \in A_2} w_j^{(k)} x_{jg} q_j} \right) w_i^{(k)} x_{ig} \tag{10}$$

When $q_j=1$ (i.e. the REE case), the replicate variance estimator of this generalized estimator, based on equation 5, is the same as the REE replicate variance estimator of Rao & Shao.

Application

Within any of the seven components of the DSE which are subject to sampling error, the cluster sums (y_i) can be broken down into two components: the total prior to any adjustments made by TES (u_i), and the additional total from the TES sample (v_i). This second piece can be further subdivided into TES totals from clusters sampled with certainty, and TES totals from clusters sampled systematically. The estimator (a DEE) of one of the components is

$$\hat{T}_{y\cdot 3} = \sum_{i \in A_2} \alpha_i u_i + \sum_{k=1}^{2} \sum_{i \in A_2} \alpha_i t_k s_{ik} a_i v_i \tag{11}$$

where s_{ik} is the third phase stratum indicator ($s_{i1}=1$ if the cluster is selected with certainty, 0 otherwise; $s_{i2}=1-s_{i1}$, an indicator that the cluster is eligible to be selected systematically), a_i is the third phase sample indicator ($a_i=1$ if the cluster is in A_3, 0 otherwise), and t_k, the TES conditional weight, is equal to

$$t_k = \frac{\sum_{i \in A_2} s_{ik}}{\sum_{i \in A_2} s_{ik} a_i} = \frac{\sum_{i \in A_2} s_{ik}}{\sum_{i \in A_3} s_{ik}} = \frac{\text{number of clusers selected in phase 2}}{\text{number of clusters selected in phase 3}} \tag{12}$$

For s_{i1}, the certainty stratum, all clusters within it have $a_i=1$, so $t_k=1$ for all clusters in the stratum.

To create the replicate estimator, we simply apply what we have learned above equations 8 and 10.

$$\begin{aligned}\hat{T}^{(j)}_{y3} &= \sum_{i\in A_2} \alpha_i^{(j)} u_i + \sum_{k=1}^{2}\sum_{i\in A_2} \alpha_i^{(j)} t_k^{(j)} s_{ik}\, a_i\, v_i \\ &= \sum_{i\in A_2} \alpha_i^{(j)} u_i + \sum_{i\in A_2} \alpha_i^{(j)} t_1^{(j)} s_{i1}\, a_i v_i + \sum_{i\in A_2} \alpha_i^{(j)} t_2^{(j)} s_{i2}\, a_i v_i\end{aligned}$$

where,

$$t_1^{(j)} \equiv 1 \qquad (13)$$

$$t_2^{(j)} = \frac{\sum_{i\in A_2} \alpha_i^{(j)} s_{i2}\, \alpha_i^{-1}}{\sum_{i\in A_2} \alpha_i^{(j)} s_{i2}\, a_i\, \alpha_i^{-1}}$$

Implementation

The first step in implementing this variance estimation methodology is calculating the replicate weights. To this point, the method of replication used to arrive at the variance is immaterial, but we will now state that we want to use the jackknife. Let the replicate weights after the first stage of sampling be the standard jackknife replicate weights

$$w_i^{(j)} = \begin{cases} 0 \text{ if } i = j \\ \dfrac{n_h}{n_h - 1} w_{hi} \text{ if i and j are in the same first phase stratum} \\ w_{hi} \text{ otherwise} \end{cases} \qquad (14)$$

Then, applying equation 10, we obtain our final replicate weights. For cluster i and replicate j:

$$\alpha_i^{(j)} = \begin{cases} 0 \;\text{ if } i=j \\ \dfrac{r_{2,i}}{r_{2,i}-1}\,\dfrac{n_{2,i}-1}{n_{2,i}}\,\dfrac{n_{1,i}}{n_{1,i}-1}\,\alpha_i \;\text{ if } i\neq j,\ i_1=j_1,\ i_2=j_2, \text{ selected in second phase} \\ \dfrac{n_{2,i}-1}{n_{2,i}}\,\dfrac{n_{1,i}}{n_{1,i}-1}\,\alpha_i \;\text{ if } i\neq j,\ i_1=j_1,\ i_2=j_2, \text{ not selected in second phase} \\ \dfrac{n_{1,i}}{n_{1,i}-1}\,\alpha_i \;\text{ if } i\neq j,\ i_1=j_1,\ i_2\neq j_2 \\ \alpha_i \;\text{ if } i\neq j,\ i_1\neq j_1 \end{cases} \qquad (15)$$

where:

i = cluster index, for the 11,303 clusters in the second phase (for the United States)

j = replicate index, based on the 29,136 clusters in the first phase

$n_{1,i}$ = the number of clusters in the same first phase stratum as cluster i

$n_{2.i}$ = the number of clusters in the same second phase stratum as cluster i
$r_{2.i}$ = the number of clusters in the same second phase stratum as cluster i and that were selected in the second phase sample
i_1, j_1 = the first phase stratum of cluster i/j
i_2, j_2 = the second phase stratum of cluster i/j

Note that this is an unusual form of the jackknife. Normally, the jackknife has as many replicates as observations. Here, we have 11,303 clusters remaining after the second phase of the sample. However, we must use replicates equal in number to the original first phase sample size, approximately 29,000. The clusters sampled out in the second phase obviously do not contribute to the variance due to the second and third phases, but they must be included to accurately account for the first phase of sampling. "Deleting" a cluster that was sampled out changes the weights of the other clusters that were in the same first phase sampling stratum.

The second step of the implementation is to adjust the imputation of certain probabilities to account for the replication. This is a component of the variance that can be accounted for by including the effect of the replicate weights in the imputation. For some persons, their match, residence, or correct enumeration status remains unresolved even after follow-up operations. In these cases, a probability for each unresolved status is imputed using an imputation cell technique, with each unresolved case in an imputation cell getting the same imputed probability. Thè general form for the imputation of the probability for an unresolved person in imputation cell k is:

$$\mathrm{Pr}_k^* = \frac{\sum_p w_p^* t_p^* \mathrm{Pr}_p}{\sum_p w_p^* t_p^*} \qquad \mathbf{(16)}$$

where the summation is over all resolved persons in imputation cell k, and:

w_p^* = person-level weight, incorporating the first and second phase of sampling, but not including noninterview adjustment

$$t_p^* = \begin{cases} \text{conditional TES weight, the inverse of the probability of selection in the TES} \\ \quad \text{sample, if the person is a TES person} \\ \text{1 if the person is NOT a TES person} \end{cases}$$

$$\mathrm{Pr}_p = \begin{cases} \text{1 if a person is a \{match/resident/correct enumeration\}} \\ \text{0 if a person is NOT a \{match/resident/correct enumeration\}} \end{cases}$$

To incorporate the replication weight, we recalculate the imputed probabilities separately for each replicate, j, in imputation cell k as:

$$\Pr{}_k^{*(j)} = \frac{\sum_p RF_p^{(j)} w_p^* t_p^{*(j)} \Pr{}_p}{\sum_p RF_p^{(j)} w_p^* t_p^{*(j)}} \tag{17}$$

where RF, the Replicate Factor, is:

$$RF_p^{(j)} = \alpha_p^{(j)} / \alpha_p$$

and

$$t_p^{*(j)} = \begin{cases} t_2^{(j)} \text{ if the person is a TES person in a non-certainty, selected cluster} \\ t_p^* \text{ otherwise} \end{cases}$$

To complete the estimation of the variances, we must compute the 29,000 replicate dual system estimates for each of the approximately 448 post-strata:

$$D\hat{S}E_h^{(j)} = (C - II)\left(\frac{CE^{(j)}}{N_e^{(j)}}\right)\left(\frac{N_n^{(j)} + N_i^{(j)}}{M_n^{(j)} + \left(\frac{M_o^{(j)}}{N_o^{(j)}}\right) N_i^{(j)}}\right) \tag{18}$$

Formula 13 must be used to separately compute each of the seven replicated terms of the DSE. We note that equation 13 operates at the cluster level, and the appropriate person-level records must be summed up to the cluster level. For convenience, three indicator variables were created to translate the three components of equation 13 into components based on person-level characteristics:

x_{ip} = 1 if the person is NOT a TES person (TESPER=0), 0 otherwise
y_{ip} = 1 if the person IS a TES person AND is from a cluster sampled with certainty in TES, 0 otherwise
z_{ip} = 1 if the person IS a TES person AND is from a non-certainty sampled cluster, 0 otherwise

For stratum h, replicate j, the dth term of the DSE is:

$$\begin{aligned}\hat{T}^{(j)}_{h,d} &= \sum_i \alpha_i^{(j)} \left(\sum_{p\in i,h} \frac{fw_{ipd}}{\alpha_i} x_{ip} \right) + \sum_i \alpha_i^{(j)} \left(\sum_{p\in i,h} \frac{fw_{ipd}}{\alpha_i} y_{ip} \right) + \sum_i \alpha_i^{(j)} t_2^{(j)} \left(\sum_{p\in i,h} \frac{fw_{ipd}}{\alpha_i} z_{ip} \right) \\ &= \sum_i RF_i^{(j)} \left(\sum_{p\in i,h} fw_{ipd}\, x_{ip} \right) + \sum_i RF_i^{(j)} \left(\sum_{p\in i,h} fw_{ipd}\, y_{ip} \right) + \sum_i RF_i^{(j)} t_2^{(j)} \left(\sum_{p\in i,h} fw_{ipd}\, z_{ip} \right)\end{aligned} \qquad \textbf{(19)}$$

where α_i is the cluster's weight after small block subsampling is complete, and fw_{ipd} is the final person-level weight incorporating whatever other probabilities and factors are needed to compute the term (e.g. match, residence, and correct enumeration probabilities, or noninterview adjustment factors).

To compute the variance estimate for stratum h:

$$Var(D\hat{S}E_h) = \sum_j \frac{n_{1,i}-1}{n_{1,i}} (D\hat{S}E_h^{(j)} - D\hat{S}E_h)^2 \qquad \textbf{(20)}$$

And finally, the variance of the national adjusted population estimate is:

$$\begin{aligned}&Var(\hat{T}_{US}) = \sum_h \sum_{h'} Cov(D\hat{S}E_h, D\hat{S}E_{h'}), \text{ where,} \\ &Cov(D\hat{S}E_h, D\hat{S}E_h) = Var(D\hat{S}E_h), \text{ and} \\ &Cov(D\hat{S}E_h, D\hat{S}E_{h'}) = \sum_j \frac{n_{1,i}-1}{n_{1,i}} (D\hat{S}E_h^{(j)} - D\hat{S}E_h)(D\hat{S}E_{h'}^{(j)} - D\hat{S}E_{h'})\end{aligned} \qquad \textbf{(21)}$$

References

1. Kim, J.K., Navarro, A., and Fuller, W., "Replication Variance Estimation for Multi-Phase Stratified Sampling", Unpublished Census Bureau Memorandum, July 24, 2000.

2. Starsinic, M., and Kim, J.K., DSSD Census 2000 Procedures and Operations Memorandum Series V-1, "Computer Specifications for Variance Estimation for Census 2000".